Supplementary Document

We conduct exhaustive qualitative and quantitative evaluation to assess our proposed fusing method.

1 Quantitative Evaluation

In this section, we show three quantitative results. First, we evaluate the semantic-level classification accuracy [3], which is shown in Table 1. Given the audio embeddings from our pre-trained audio encoder, we train a linear classifier to recognize eight semantic labels including giggling, sobbing, nose-blowing, wind, fire crackling, underwater bubbling, explosion, and thunderstorm. Second, we measure cosine similarity between text-guided and sound-guided latent code. We notice that two different domain latent codes are pointing similar direction in the embedding space (see Table 2). Finally, we apply our method on zero-shot task to demonstrate the usefulness of our approach. We evaluate the distinguishability of the feature vector from the proposed audio encoder by comparing the downstream zero-shot classification task. As a baseline model, we use a ResNet50-based classifier [1], which is trained end-to-end from scratch (i.e. random initialization). From the experimental results, our method outperforms the ResNet50 model as shown in Table 3.

Table 1: Semantic-level classification accuracy.

	Attribute ($\uparrow$)							
Modal	Giggling	Sobbing	Nose blowing	Wind	Fire crackling	Underwater bubbling	Explosion	Thunderstorm
Text	0.89	0.67	0.74	0.89	1.00	0.96	1.00	1.00
Audio	1.00	0.97	0.89	1.00	0.88	1.00	0.99	0.99

Table 2: Cosine similarity between text-guided and sound-guided latent code.

	Attribute ($\uparrow$)							
Name	Giggling	Sobbing	Nose blowing	Wind	Fire crackling	Underwater bubbling	Explosion	Thunderstorm
Mean	0.996	0.993	0.997	0.759	0.759	0.759	0.758	0.761
Std	0.002	0.001	0.002	0.005	0.006	0.005	0.006	0.005

Table 3: CLIP-based audio encoder zero-shot inference accuracy.

	Dataset ($\uparrow$)	
Model	ESC-50	Urban sound 8k
ResNet50 for audio classification [1]	0.668	0.713
CLIP-based audio encoder	0.622	0.731

2 Qualitative Evaluation

In this section, we show two qualitative results to show the effectiveness of using audio mixture model for the image manipulation. First, we compare the quality of our method to the text-based image manipulation method. As shown in Figure 1, proposed method contains more vivid content than the text-based method. Also, we perform mixture of content from the text and the style from the audio, which is novel method that fuses text and audio information for novel image generation (see Figure 1). We show that the model is not memorizing the dataset but actually learning the meaningful smooth embedding space, we perform latent code interpolation between two latent codes from distinct attributes. Generated images show the smooth changes along two different attributes as shown in Figure 2.

1st CtrlGen: Controllable Generative Modeling in Language and Vision Workshop at NeurIPS 2021.

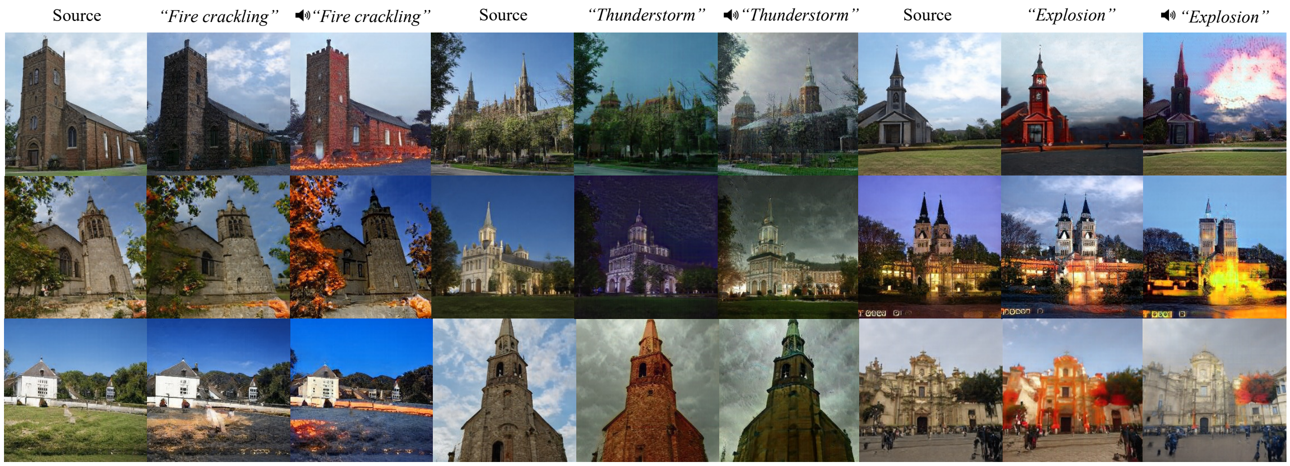
References

[1] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[2] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[3] A. Voynov and A. Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pages 9786–9796. PMLR, 2020.

[4] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

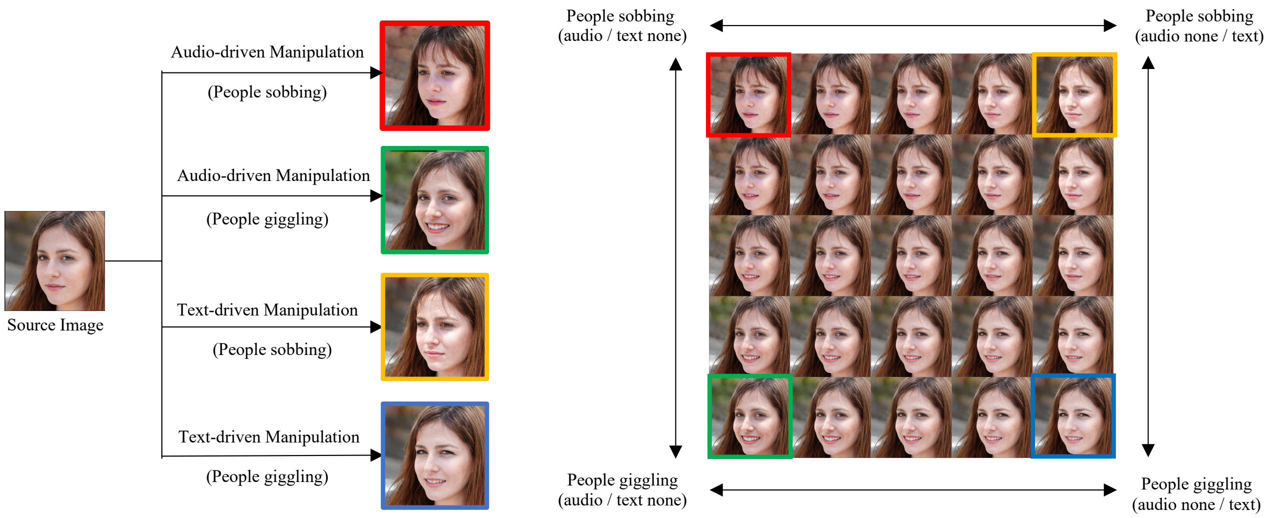


(a) Results of text-driven manipulation and audio-driven manipulation from LSUN dataset [4].

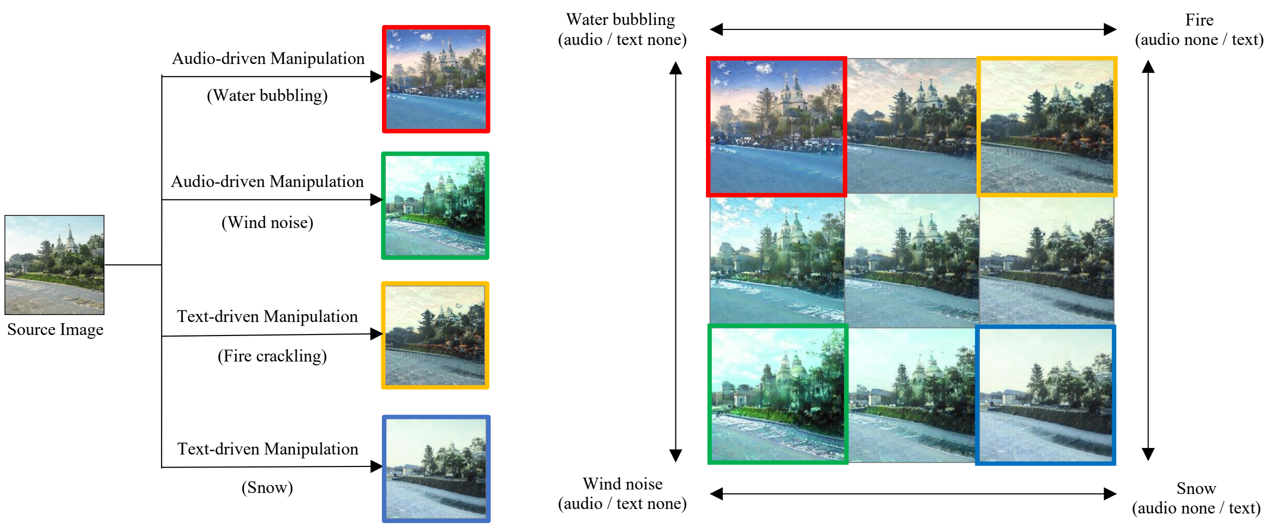
(b) Results of text-driven manipulation and audio-driven manipulation from FFHQ dataset [2].

(c) The result of style mixing between the sound-guided fine-grained style and text-driven high-level style.

Figure 1: Results of text-driven manipulation and audio-driven manipulation. (a)-(b) shows the results of manipulation by source, text-driven latent optimization, and sound-guided latent optimization for each attribute. The 1st, 4th, and 7th columns are the source image, the 2nd, 5th, and 8th columns are the results of guiding the image with text, and the 3rd, 6th, and 9th columns are the results of guiding the image with audio. (c) shows the result of style-mixing even when the latent code is optimized with different modal.



(a) In the latent space of pre-trained StyleGAN2 with FFHQ, the latent code is guided by text and audio modal for the attribute of "people giggling, people sobbing," and the interpolation result between the changed latent codes is shown.



(b) It shows that interpolation between latent codes is possible even if latent codes are guided for different modals and attributes in the latent space of StyleGAN2 pre-trained with LSUN (church).

Figure 2: Interpolation result of the optimized latent code. Since the audio embedding is mapped to the CLIP space, the latent code obtained by guiding the latent code of the source image to various modals can be interpolated by attribute or even by modal.